Introducción a las MLOps y el Ecosistema MLFlow 3.X

Aprendizaje Automático Aplicado

Julio Waissman Vilanova 2026

Maestría en Ciencia de Datos / Universidad de Sonora

¿Y que es eso de MLOps?

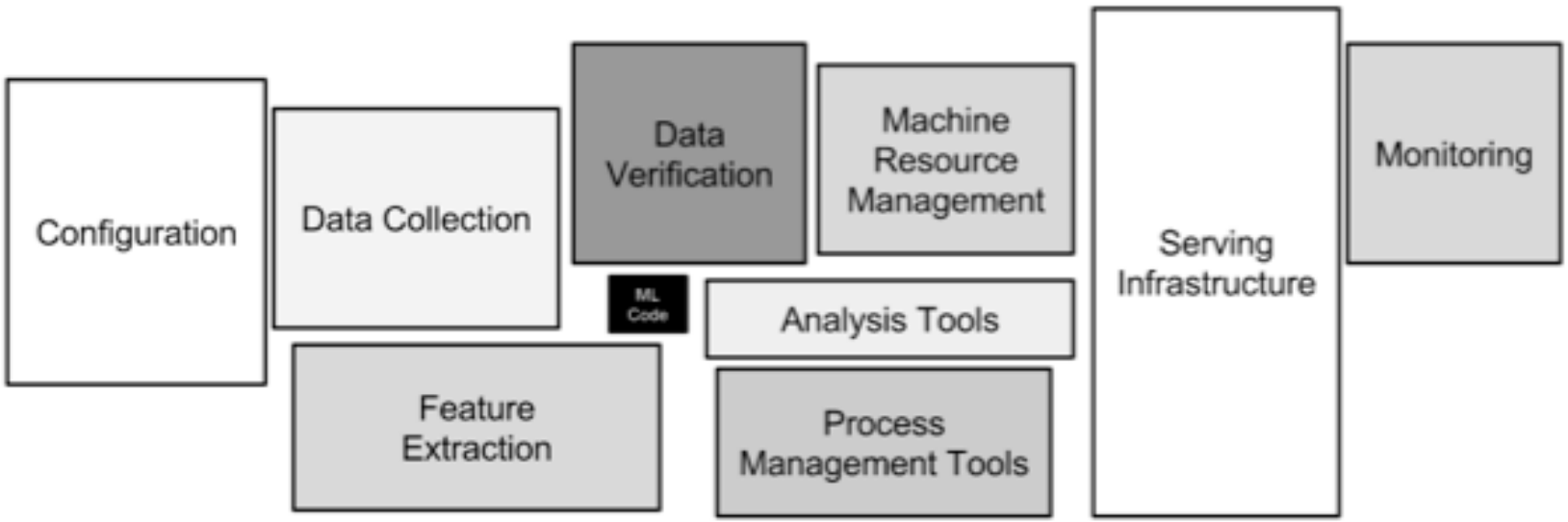
El Problema: El *Valle de la Muerte* del ML

Muchos modelos de ML mueren en la fase de prototipo. ¿Por qué?

- **Código vs. Sistema:** El código del modelo es solo una pequeña fracción del sistema total.
- **Deuda Técnica:** Falta de pruebas, versionado y reproducibilidad.
- **Desconexión de equipos:** Científicos de datos (experimentación) vs. Ingenieros de software (producción).

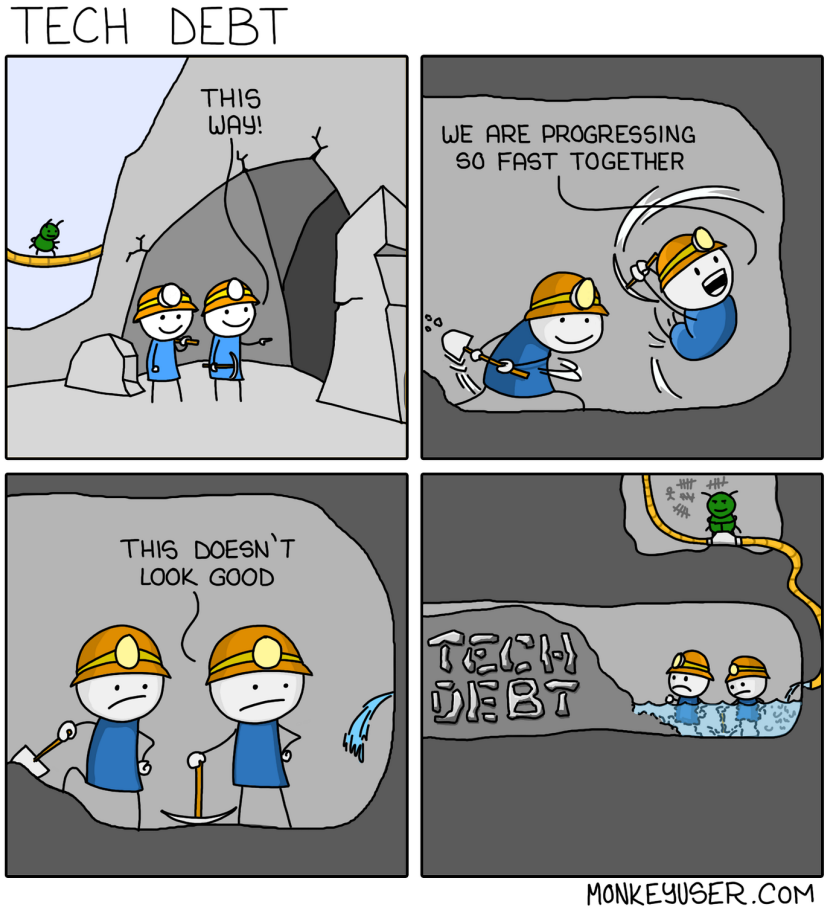
Código vs Sistema

Los sistemas con ML suelen ser complejos



Deuda técnica

Machine Learning «The high-interest credit card of technical debt»



Deuda técnica

ML vs Software tradicional

	ML	**Tradicional**
Dependencias	A los datos y al código	Al código
Validez	Degrada en el tiempo	Lógica estable
Comlejidad	Varios flujos de trabajo interconectados	Procesos lineales
Herramientas	Evolucionan y cambian muy rápido	Relativamente estable
Transparencia	Limitada (en particular con modelos preentrenados)	Flujo lógico claro

¿Qué es MLOps?

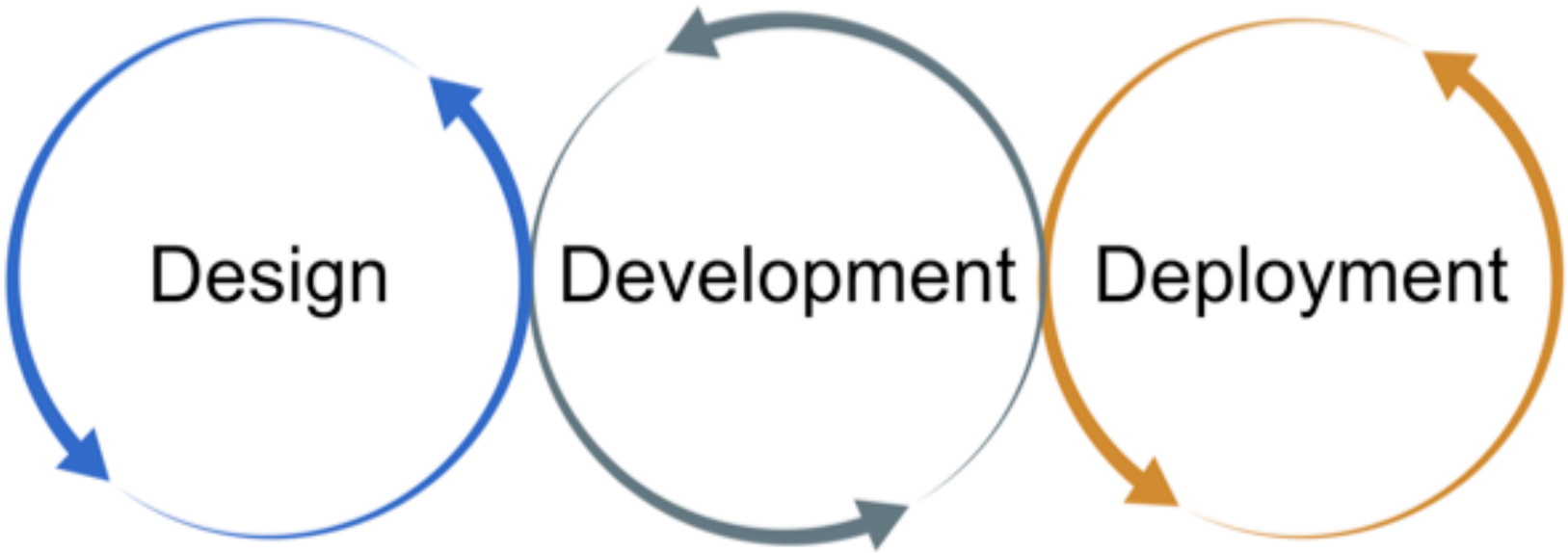
MLOps (Machine Learning Operations) es una cultura y práctica de ingeniería que busca unificar el desarrollo de sistemas de ML (Dev) y la operación de sistemas de ML (Ops).

Objetivo Principal
Estandarizar y automatizar el ciclo de vida de los modelos de aprendizaje automático.

- **CI (Integración Continua):** Pruebas de código, datos y validación de modelos.
- **CD (Entrega Continua):** Despliegue automático en servicios de predicción.
- **CT (Entrenamiento Continuo):** Re-entrenamiento automático basado en nuevos datos.

El ciclo de vida de MLOps

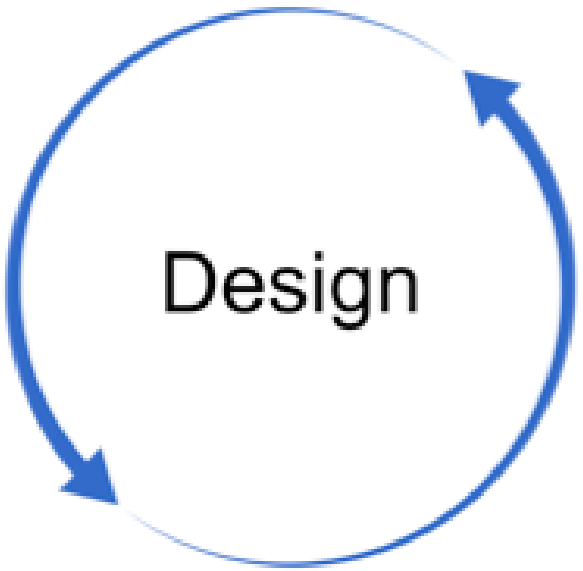
El ciclo de vida completo



Tres etapas iterativas e interconectadas. Es normal ir y venir entre etapas

Diseño

- Comprensión del negocio
- Comprensión de los datos
- Diseño de solución de ML



Desarrollo (ML)

- Ingeniería de datos
- Ingeniería de características
- Desarrollo de modelos
- Prueba de concepto (PoC)

Despliegue (Ops)

- Pruebas
- Despliegue
- Monitoreo
- Reentrenamiento y mejora continua

Operaciones Esenciales de MLOps

Ingeniería de Datos

Sin datos de calidad, no hay modelo de calidad.

- **Ingesta:** Captura de fuentes diversas
- **Etiquetado:** Posiblemente la etapa más costosa (tanto en tiempo como en dinero)
- **Versionado de datos:** Registrar modificaciones en los datos
- **Validación de datos:** Detección de anomalías en los tipos de datos o distribuciones antes del entrenamiento.
- **Feature Store:** Almacén centralizado de características para entrenamiento y despliegue.

Desarrollo de modelos

El entrenamiento debe ser reproducible y rastreable.

- **Versionado de experimentos:** Registrar hiperparámetros, métricas y artefactos.
- **Orquestadores:** Uso de herramientas para gestionar las dependencias de las tareas de entrenamiento.
- **Evaluación de modelos:** Comparación contra baselines y pruebas de sesgo/equidad.

Despliegue

El trabajo no termina cuando el modelo se despliega.

- **Estrategias de despliegue:** Canary deployment, A/B Testing, Shadow deployments.
- **Versionado de Modelos:** Versones de modelos en desuso, de prueba y en operación.
- **Monitoreo:**
  - *Data Drift:* Cambio en la distribución de entrada.
  - *Concept Drift:* Cambio en la relación entre entrada y salida.
- **Retroalimentación:** Capturar predicciones reales para futuros re-entrenamientos.

Introducción a MLFlow 3.X

¿Qué es MLFlow 3.X?

MLFlow es la plataforma de código abierto más utilizada para gestionar el ciclo de vida de ML de extremo a extremo.

Novedades de la versión 3.X:

- **IA Generativa (LLMOps):** Soporte nativo para el rastreo de prompts, evaluación de LLMs y despliegue de agentes.
- **Escalabilidad:** Mejoras en la gestión de bases de datos de metadatos a gran escala.
- **Interfaz Unificada:** Experiencia mejorada para comparar miles de ejecuciones (runs).

Arquitectura de MLFlow 3.X

Se divide en componentes modulares que pueden usarse de forma independiente o conjunta:

1. MLFlow Experiment Tracking
2. MLFlow Datasets
3. MLFlow Models
4. MLFlow Evaluate
5. MLFlow Model Registry
6. MLFlow Serving
7. MLFlow Projects

MLFlow Experiment Tracking

Es el núcleo de la experimentación. Registra y permite consultar:

- **Parámetros:** Diccionario de entradas (ej: `learning_rate`).
- **Métricas:** Valores numéricos para seguimiento (ej: `Accuracy`, `F1-score`).
- **Artefactos:** Archivos de salida (gráficos, pesos del modelo, archivos de configuración).
- **Tracing (3.X):** Visualización paso a paso de cadenas de LLM (LangChain, LlamaIndex).

MLFlow Model

Define un formato estándar para empaquetar modelos.

- **Sabor (Flavor):** Soporte para ‘scikit-learn‘, ‘pyTorch‘, ‘TensorFlow‘, etc.
- **Estructura:** Una estructura compleja que incluye dependencias, metadatos, firmas, ...
- **Flexibilidad:** Extiene su uso con los modelos de la clase ‘PyFunc‘.

MLFlow Model Registry

Un repositorio central de modelos para gestionar versiones y estados.

- **Versionado:** Registro automático de nuevas iteraciones.
- **Trazabilidad:** Todos los modelos pueden trazarse desde Experiment Tracking, e inclusive con la firma de los datos utilizados para entrenamiento.
- **Flujos de producción simplificados:** Transiciones entre *Campeón*, *Retador*, y *Archivado* entre otros. Fácil de regresar a modelos anteriores.
- **Gobernanza:** Control de quién puede promover un modelo a producción mediante flujos de aprobación.

MLFlow Serving

- **Despliegue:** Los modelos pueden servirse como:
  - REST API local.
  - Contenedores en AWS SageMaker, Azure ML o Google Vertex AI.
  - Funciones personalizadas en Python.
- **Despliegue simplificado** interface general que simplifica el proceso de puesta en producción.
- **Estandarizado y abierto:** Evita el *vendor lock-in*.

Lo nuevo en la versión 3.X.

- **Datasets:** Manejo de firmas y uso de datos de entrenamiento y validación, que permite un mejor seguimiento de la repetibilidad de los experimentos.
- **Evaluate:** Un sistema para evaluar modelos de aprendizaje en forma estructurada. Incluye un modulo SHAP para un intento básico (por el momento) de XAI.
- **AI Gateway:** Un proxy seguro para gestionar credenciales y límites de tasa para proveedores de LLM (OpenAI, Anthropic, etc.).
- **Evaluation:** Herramientas para comparar prompts y modelos generativos usando métricas como *Perplexity* o evaluadores basados en otros LLMs.

¿Preguntas?

Vamos entonces a ver como funciona MLFlow en local y en linea.

https://mlflow.org